基于 K-means 和 naive Bayes 的数据库用户行为异常检测研究 *

王旭仁 [1,2]，冯安然 [1,2†]，何发镁 [2,3]，马慧珍 [1,2]，杨　杰 [1,2]

(1. 首都师范大学 信息工程学院，北京 100048; 2. 中国科学院信息工程研究所 中国科学院网络测评技术重点实验室，北京 100093; 3. 北京理工大学 图书馆，北京 100081)

摘　要：针对数据库用户行为异常导致数据库泄露问题，提出了一种基于 K-means 和 naive Bayes 算法的数据库用户异常检测方法。首先，利用数据库历史审计日志中用户的查询语句与查询结果，采用 K-means 聚类方法得到用户的分组；然后，使用 naive Bayes 分类算法构造用户异常检测模型。与单独使用 naive Bayes 分类法构造的模型相比，在数据预处理时精简了用户行为轮廓的表示方法，降低了计算冗余，减少了 81%的训练时间；利用 K-means 聚类方法得到用户组别，使检测的精确率提高了 7.06%，F1 值提高了 3.33%。实验证明，所提方法大幅降低训练时间，取得了良好的检测效果。

关键词：数据库；用户行为；异常检测；K-means 聚类；naive Bayes 分类算法

中图分类号：TP393.08　　doi: 10.19734/j.issn.1001-3695.2018.09.0755

Research of database user behavior anomaly detection based on K-means and naive Bayes

Wang Xuren[1, 2], Feng Anran[1, 2†], He Famei[2, 3], Ma Huizhen[1, 2], Yang Jie[1, 2]

(1. *Information Engineering College*, *Capital Normal University*, *Beijing* 100048, *China*; 2. *Key Laboratory of Network Assessment Technology*, *Institute of Information Engineering*, *Chinese Academy of Sciences*, *Beijing* 100093, *China*; 3. *Library*, *Beijing Institute of Technology*, *Beijing* 100081, *China*)

Abstract: Aiming at database leakage caused by abnormal database user behavior, this paper proposed a database user anomaly detection method based on K-means and Naive Bayes algorithm. Firstly, the K-means clustering method obtained users' grouping based on the user's query statements and query results in the database historical audit logs; then, the Naive Bayes classification algorithm constructed the user anomaly detection model. Compared with the model constructed by Naive Bayes classification alone, the simplified representation of user behavior profile reduces computational redundancy and reduces training time by 81%. Applying K-means clustering method to obtaining users' grouping improves the detection accuracy by 7.06% and the F1 value by 3.33%. Experiments show that the proposed method greatly reduces the training time and achieves better detection results.

Key words: database; user behavior; anomaly detection; K-means clustering; naive Bayes classification

0　引言

Anderson[1]在 1980 年首次提出了入侵检测的概念，引发了入侵检测系统的研究。入侵检测技术分为异常检测和误用检测：异常检测技术先定义“正常情况”下的观测数据，通过对比新生成的数据与正常数据的偏差得出系统是否有被攻击的迹象；误用检测技术则是通过收集异常数据将之归纳为一个模型，符合此模型的数据会被判定为异常。近年来对异常检测技术的研究较为活跃。Hu Qiaona 等人 [2,3]整合了网络中不同来源的用户数据进行特征提取构造异常检测器；Stanislav 等人[4]挖掘了网络传输中数据包的信息作为特征；Ruan Xin 等人[5,6]对社交网络中的用户行为进行风险评估。

K-means 聚类和 naive Bayes 分类在异常检测领域得到了广泛应用。Shin 等人[7]使用 K-means 聚类得到网络系统的正常状态与异常点，根据不同状态间的关系构造 Markov 概率模型对网络状态进行概率评估与预判。Louvieris 等人[8]先对网络数据使用 K-means 聚类得到异常数据的类别，再利用 naive Bayes 对特征的相关性和重要性进行排序，从中提取不同类别的特征构造 C4.5 决策树对异常数据进行分类检测。这种方法充分利用了每个分类器的特点，提高了系统异常检测的准确率。Karami 等人[9]在 K-means 聚类的基础上加入了 PSO（粒子群优化）算法，利用 PSO 的全局搜索能力找出最优初始聚类中心点，使用 K-means 聚类则可以避免 PSO 的局部最优解，二者相结合提高了系统的异常检测率。李洪成等人[10]采用了遗传 K-means 算法作为移动自组网（MANET）的异常检测方法，将遗传算法与 K-means 相结合可实现聚类结果的全局最优。Kreimel 等人[11]通过分析信息物理网络中数据的特征计算出异常值的范围，利用异常数据的特征训练 Naive Bayes 分类模型。贾凡等人[12]针对网络数据中的不同攻击类型，在各个类型相关性最大的维度使用分层 K-means 算法，通过提高对每种攻击的检测率提高了整体的异常检测率。以上这些方法主要对网络中的数据进行挖掘和分析，不能直

收稿日期：2018-09-27；**修回日期：**2018-11-12　**基金项目：**国家自然科学基金资助项目（61373161）；中国科学院信息工程研究所中国科学院网络测评技术重点实验室 2018 开放课题

作者简介：王旭仁（1972-），女，贵州毕节人，副教授，博士，主要研究方向为数据挖掘、网络信息安全；冯安然（1993-），女（通信作者），河北石家庄人，硕士研究生，主要研究方向为数据挖掘、网络信息安全（ann654175863@163.com）；何发镁（1972-），男，四川绵阳人，博士，主要研究方向为情报分析、数据挖掘；马慧珍（1990-），女，河南周口人，硕士研究生，主要研究方向为数据挖掘、网络信息安全；杨杰（1994-），男，山西晋城人，硕士研究生，主要研究方向为数据挖掘、网络信息安全.

接用于数据库的用户异常检测。

1　相关工作

为了保证数据库中数据的安全可靠和正确有效，一些学者[13~15]对数据库的访问权限控制方法进行了改进，使权限控制策略能够解决更加复杂的实际应用问题。这类方法可以抵御一部分外部攻击，但是不能检测出来自内部用户的攻击和恶意用户的伪装攻击。

Ashish 等人[16]基于企业数据库管理系统的结构，提出了一种针对数据库的异常检测模型。该文献分别研究了两种不同场景下使用的异常检测方法：在包含用户分组结构的数据库管理系统中使用 Naive Bayes 分类法来构造异常检测模型；在无用户分组结构时采用 K-means 等聚类算法先对用户进行分组，再利用离群值检测法构造异常检测模型。这篇文献提出的异常检测方法能够检测出大部分的用户异常，但是由于只考虑了用户提交查询语句的语法结构，并不能检测出用户伪装异常。

文献[17]与文献[16]中的方法作对比，将查询语句的返回结果作为用户行为特征，分别构造了 naive Bayes、决策树（decision tree）、支持向量机（SVM）三种分类器。实验表明将查询结果作为用户行为特征可以有效地检测出用户的伪装异常。但是，由于查询结果往往数目比较庞大，利用其统计特征来构造用户模型花费的时间较长，训练效率较低。

Asmaa 等人 [18]将查询语句的语法结构和查询结果结合在一起作为用户行为特征，分别构造了 naive Bayes 与多标签分类器。这种方法并没有计算返回结果的各项统计特征，而是将结果的数目占总查询表的比例作为一项特征加入到用户行为轮廓中，这样既考虑了查询的语义语法，又没有增加过多的计算量，检测效果也有所提升。但是，在使用仿真数据集时没有根据具体的数据库查询语句修改用户轮廓的构建方法，计算空间存在冗余，增加了计算成本；此外，在对用户进行分组时，采用了定义的方法，不能准确地对用户的行为进行分组，可能会造成一定的误差。

针对文献[18]检测方法的不足，本文提出一种改进方法——将构成用户轮廓的向量表示方法精简为针对该数据库查询方式的结构，减少计算冗余；同时，使用 K-means 聚类的方法对用户进行分组，使用户的组别更加符合其行为特征，提高了整体的检测率。

2　相关算法

2.1　K-means 算法

假设数据集 D 包含 n 个欧氏空间的对象。划分方法把 D 中的对象分配到 K 个簇 $C_1,\ldots,C_K$ 中，使得对于 $1\le i, j\le K$, $C_i\subset D$ 且 $C_i\cap C_j=\varnothing$。用一个目标函数来评估划分质量，使得簇内对象相互相似，而与其他簇内的对象相异。也就是说，该目标函数以簇内高相似性和簇间低相似性为目标[19]。

基于形心的划分技术使用簇 c_i 的形心代表该簇。从概念上讲，簇的形心是它的中心点。对象 $p\in C_i$ 与该簇的代表 c_i 之差可以用 $\mathrm{dist}(p,c_i)$度量，其中 $\mathrm{dist}(x,y)$是两个点 x 和 y 之间的欧氏距离。目标函数的定义如下：

$$P(Y=c_k), k=1,2,\ldots,K \tag{1}$$

其中，E 是数据集中所有对象的误差平方和；p 是空间中的点，表示给定的数据对象；c_i 是簇 C_i 的形心。

K-means 算法过程如下：

a) 从 D 中任意选择 K 个对象作为初始簇中心；

b) repeat

c)根据簇中对象的均值，将每个对象分配到最相似的簇；

d) 更新簇均值，即重新计算每个簇中对象的均值；

e) until 目标函数收敛；

K-means 聚类算法的局限性在于数据初始化，聚类中心初始值的选择会影响最终的聚类结果。对此，本文进行了多次实验，得到不同初始值情况下的聚类结果，比较各结果的均值方差，取最小方差的聚类模型作为用户分组结果。

2.2　Naive Bayes 算法

设输入空间 $\alpha\subseteq R^n$ 为 n 维向量的集合，输出空间为类标记集合 $\beta=\{c_1,c_2,\ldots c_K\}$。输入为特征向量 $x\in\alpha$，输出为类标记 $y\in\beta$。X 是定义在输入空间 α 上的随机向量，Y 是定义在输出空间 β 上的随机变量。$P(X,Y)$是 X 和 Y 的联合概率分布。训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\ldots,(x_N,y_N)\}$由 $P(X,Y)$独立同分布产生[20]。

Naive Bayes 算法通过训练数据集学习联合概率分布 $P(X,Y)$。具体的，学习以下先验概率分布及条件概率分布。先验概率分布为

$$P(Y=c_k), k=1,2,\ldots,K \tag{2}$$

条件概率分布为

$$P(X=x|Y=c_k)=P(X^{(1)}=x^{(1)},\ldots,X^{(n)}=x^{(n)}|Y=c_k) \quad k=1,2,\ldots,K \tag{3}$$

于是学习到联合概率分布 $P(X,Y)$。

Naive Bayes 算法对条件概率分布作了条件独立性假设

$$P(X=x|Y=c_k)=\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k) \tag{4}$$

算法的具体过程为：对给定的输入 x，通过学习到的模型计算后验概率分布 $P(Y=c_k|X=x)$，将后验概率最大的类作为 x 的类输出。后验概率计算根据贝叶斯定理进行：

$$P(Y=c_k|X=x)=\frac{P(X=x|Y=c_k)P(Y=c_k)}{\sum_k P(X=x|Y=c_k)P(Y=c_k)} \quad k=1,2,\ldots,K \tag{5}$$

将式(4)代入式(5)有

$$P(Y=c_k|X=x)=\frac{P(Y=c_k)\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)}{\sum_k P(Y=c_k)\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)} \quad k=1,2,\ldots,K \tag{6}$$

这是 naive Bayes 分类（NBC）的基本公式。于是，naive Bayes 分类器可表示为

$$y=f(x)=\arg\max_{c_k}P(Y=c_k|X=x) =\frac{P(Y=c_k)\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)}{\sum_k P(Y=c_k)\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k)} \tag{7}$$

3　基于 K-means 和 naive Bayes 算法的异常检测系统

3.1　系统结构

系统的整体结构如图 1 所示。系统的工作流程分为两个阶段，分别为训练和测试阶段。

训练阶段步骤如下：

a) 对历史审计日志进行预处理，去除系统日志后得到用户查询数据；

b) 提取查询数据的特征，得到特征向量，即为用户的行为轮廓；

c) 利用 K-means 算法对用户的行为轮廓进行聚类，得到用户的组别；

d) 使用 Naive Bayes 分类法对训练数据进行训练，得到异常检测模型；

测试阶段步骤如下：

a) 对用户提交的查询进行预处理；

b) 提取查询数据的特征，得到特征向量；

c) 将用户的特征向量输入异常检测模型中，得到检测结果；

d) 将检测结果输入响应器中，根据预先设定的响应策略发出响应。

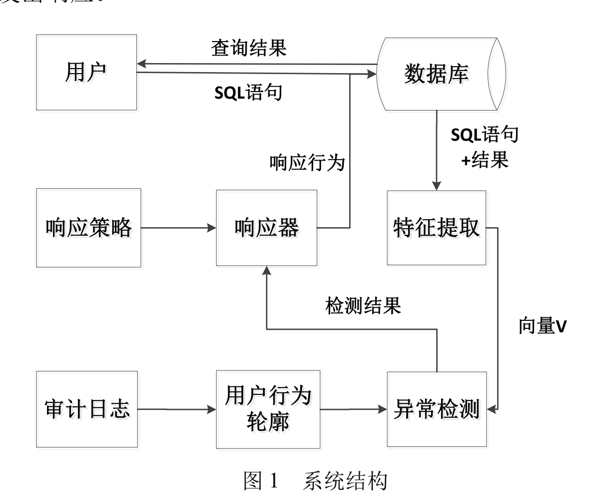


图 1　系统结构

Fig. 1　System architecture

3.2　数据表示

使用一个向量 $V(C,T,A,R)$来表示一条用户提交的查询。其中，C 为语句的命令类型，T 为查询的表格，用 0—N 表示。若数据库中仅包含少量的交叉查询，为了减少向量的存储空间，将几种交叉查询检索的表中的属性重新整合为几个新表，从 N+1 开始表示。A 为查询语句检索的属性信息，由一个数组来表示，数组的长度为所有表长度中的最大值；当查询中包含表中的某个属性时，将该属性所在位设为 1，否则为 0。R 为查询结果所占的比例，计算方式为查询结果的行数与表的总行数之比。V 向量的示例如表 3 中的第三列所示。第一列为查询语句的举例，其中用户表（client）如表 1 所示，产品表（products）如表 2 所示。第二列为 Q 向量[18]，分别包含四个属性：命令类型（C）、检索表向量（Pr），检索属性（Pa），检索信息的比例（Sr）。使用 V 向量表示交叉查询较少的用户查询，与向量 Q 相比节省了一半的存储空间，可以有效地降低异常检测模型的训练时间。

表 1　用户表

Table 1　Clients' table

c_ID	c_name
1	c1
2	c2
3	c3
4	c4

表 2　产品表

Table 2　Products' table

p_ID	price
1	1
2	2
3	5
4	8

表 3　向量表示

Table 3　Vector representation

Query	Q(C,Pr,Pa,Sr)	V(C,T,A,R)
SELECT * FROM Clients WHERE c_ID=3;	('SELECT', [1,0], [[0,0,1,0],[0,0,0,0]], ['s',null])	('SELECT', 0, [0,0,1,0], 0.25)
SELECT * FROM Products WHERE price<5;	('SELECT', [0,1], [[0,0,0,0],[1,1,0,0]], [null,'m'])	('SELECT', 1, [1,1,0,0], 0.5)

4　实验结果与分析

4.1　数据集及评估方法

由于真实数据库的后台审计日志不易获取，本次实验采用 TPC-C 数据库作为实验数据集。TPC(Transaction Processing Performance Council) 事务处理性能协会是一个评价大型数据库系统软硬件性能的非盈利性组织。TPC 制定的规范在数据库异常检测领域已有多次应用[18,21,22]。TPC-C[23] 是 TPC 协会制定的专门针对联机交易处理系统（OLTP 系统）的规范。TPC-C 测试用到的模型是一个大型的商品批发销售公司，它拥有若干个分布在不同区域的商品仓库。该系统需要处理的交易事务主要分为以下五种：新订单、支付操作、发货、订单状态查询、库存状态查询等。

为了对实验结果进行评估，引入以下三个评价指标：

精确率(precision)：反映了被分类器判定的正常样本中真正的正常样本的比重，其定义如下：

$$P = \frac{TP}{TP + FP} \tag{8}$$

召回率(recall)，也称为 true positive rate，反映了被正确判定的正常样本占总的正常样本的比重，定义如下：

$$R = \frac{TP}{TP + FN} \tag{9}$$

F1 值：模型精确率和召回率的一种加权平均，定义如下：

$$F = \frac{2PR}{P + R} \tag{10}$$

其中：

TP—将正常样本预测为正常类数；

FN—将正常样本预测为异常类数；

FP—将异常样本预测为正常类数；

TN—将异常样本预测为异常类数。

4.2　预处理

在 Linux 系统中，利用 tpcc-mysql 工具构建 TPC-C 数据库，创建表，模拟出商品批发销售公司的五种交易事务。tpcc-mysql 是由 percona 公司基于 TPC-C 衍生出来的产品，专门用于构建 TPC-C 的标准数据库。交易模拟完成后，利用 Mysql 的日志功能得到 91121 条审计日志。去除审计日志中的系统命令等与用户交易无关的数据得到 86924 条用户查询。使用 3.2 小节中的数据表示方法将用户的查询转换为 V 向量，作为用户行为的特征向量。

4.3　实验结果

使用 K-means 算法对经过预处理的用户特征向量进行聚类得到用户的组别。根据 TPC-C 数据库模拟的场景，将 k 设为 2，分别代表数据库中的两个组别——客户与工作人员。

由于 K-means 算法的局限性，改变初始簇中心对 K-means 算法进行多次实验。如图 2 所示，不同的随机种子数目会产生不同的初始类簇中心点，导致最后的聚类结果发生变化。由图可知最小的平方误差和为 108849，因此选取此

聚类模型作为用户分组的依据。

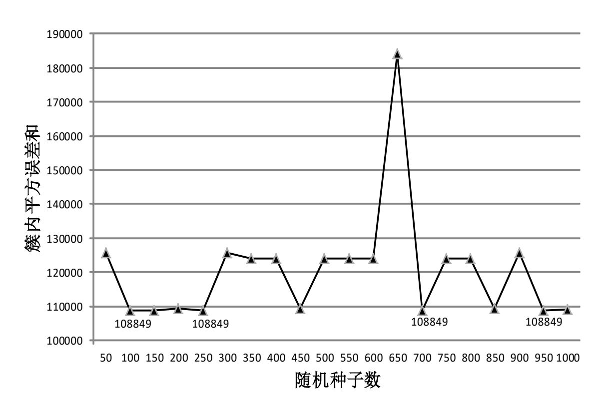


图 2 初始聚类中心点对聚类效果的影响

Fig. 2 Influence of initial cluster center points on clustering effect

K-means 聚类算法对用户进行分组后，将用户组别作为标签，使用 naive Bayes 分类算法训练分类器。

训练时间与训练数据量如表 4 所示。本文实验使用的数据量是文献[18]中数据集的 9.39 倍，而训练时间则是 1.78 倍。去除了向量的冗余之后，训练时间降低了 81%。

表 4 分类器的训练时间与训练数据量

Table 4 Training time of classifiers and training data size

	数据量（条记录）	训练时间/s
K-means+NBC	69.5k	1.51
NBC[18]	7.4k	0.85

如表 5 所示，将本实验的结果与文献[18]中单独使用 naive Bayes 分类算法的结果作对比，发现召回率降低了 0.72%，而精确率提升了 7.06%，总体来看，F1 值提升了 3.33%。精确率的提升表明，使用 K-means 聚类使得用户的分组更加符合其行为特征，因此对异常用户的识别能力有所增强；而召回率的降低表明，精简的用户行为特征向量虽然降低了 81%的训练时间，却弱化了向量对正常用户特征行为的表征能力，因此对正常用户的识别有所降低。作为两项评估标准的加权平均，F1 值提高了 3.33%，且本次实验使用的数据量是文献[18]中的 9.39 倍，这表明在增大数据量的情况下，本文使用的方法检测效果更好，鲁棒性更高。

表 5 分类器结果对比

Table 5 Result comparison of classifiers

	precision	recall	F1-score
K-means+NBC	97.17	97.16	97.16
NBC[18]	90.11	97.88	93.83

5 结束语

数据库泄露问题日益严峻。为了减少数据库泄露事件的发生，本文提出了一种基于 K-means 聚类和 naive Bayes 分类的用户异常检测模型。实验证明，本文的模型训练时间短，精确度高，具有一定的鲁棒性。在接下来的研究中，将继续探索其他的模型构造方法，以期获得更好的检测效果。

参考文献：

[1] James P A. Computer security threat monitoring and surveillance, TR80904 [R].Washington: James P A Co, 1980.

[2] Hu Qiaona, Tang Baoming, Lin D. Anomalous user activity detection in enterprise multi-source logs [C]//Proc of the 17th IEEE International Conference on Data Mining. Washington DC: IEEE Computer Society, 2018: 797-803.

[3] Sreyasee D B, Yuan Junsong, Zhang Jiaqi, *et al.* Context-aware graph-based analysis for detecting anomalous activities [C]//Proc of IEEE International Conference on Multimedia and Expo. Washington DC: IEEE Computer Society, 2017: 1021-1026.

[4] Stanislav P, Travis A. Industrial control system network intrusion detection by telemetry analysis [J]. IEEE Trans on Dependable and Secure Computing, 2016, 13 (2): 252-260.

[5] Ruan Xin, Wu Zhenyu, Wang Haining, *et al.* Profiling online social behaviors for compromised account detection [J]. IEEE Trans on Information Forensics and Security, 2016, 11 (1): 176-187.

[6] Laleh N, Barbara C, Ferrari E. Risk assessment in social networks based on user anomalous behaviors [J]. IEEE Trans on Dependable and Secure Computing, 2018, 15 (2): 295-308.

[7] Shin S, Lee S, Kim H, *et al.* Advanced probabilistic approach for network intrusion forecasting and detection [J]. Expert Systems with Applications, 2013, 40 (1): 315-322.

[8] Louvieris P, Clewley N, Liu X. Effects-based feature identification for network intrusion detection [J]. Neurocomputing, 2013, 121 (18): 265-273.

[9] Karami A, Guerrero Z M. A fuzzy anomaly detection system based on hybrid PSO-Kmeans algorithm in content-centric networks [J]. Neurocomputing, 2015, 149: 1253-1269.

[10] 李洪成，吴晓平，严博. 面向 MANET 异常检测的分布式遗传 K-means 研究 [J]. 通信学报, 2015, 36 (11): 167-173. (Li Hongcheng, Wu Xiaoping, Yan Bo. Research on distributed genetic k-means for anomaly detection in MANET [J]. Journal on Communications, 2015, 36 (11): 167-173.)

[11] Kreimel P, Eigner O, Tavolato P. Anomaly-based detection and classification of attacks in cyber-physical systems [C]// Proc of the 12th International Conference on Availability, Reliability and Security. Reggio Calabria: Association for Computing Machinery, 2017: 1-6.

[12] 贾凡，严妍，张家琪. 基于 K-means 聚类特征消减的网络异常检测 [J]. 清华大学学报:自然科学版, 2018, 58 (2): 137-142. (Jia Fan, Yan yan, Zhang Jiaqi. K-means based feature reduction for network anomaly detection [J]. Journal of Tsinghua University:Science and Technology, 2018, 58 (2): 137-142.)

[13] Qun Ni, Alberto T, Elisa B, *et al.* Privacy-aware role-based access control [J]. ACM Trans on Information and System Security, 2010, 13 (3): 1-31.

[14] Mehdi H, Jovan S, Annamaria C. Access control for data integration in presence of data dependencies [C]//Proc of the 19th International Conference on Database Systems for Advanced Applications. Switzerland: Springer Verlag, 2014: 203-217.

[15] José R D, Víctor J R, Amir H C. Biometric access control for e-health records in pre-hospital care [C]// Proc of EDBT/ICDTJoint Conference. 2013: 169-173.

[16] Ashish K, Evimaria T, Elisa B. Detecting anomalous access patterns in relational databases [J]. VLDB Journal, 2008, 17 (5): 1063-1077.

[17] Sunu M, Michalis P, Hung N, *et al.* A data-centric approach to insider attack detection in database systems [C]//Proc of the 13th International Symposium Recent Advances in Intrusion Detection. Springer-Verlag, 2010: 382-401.

[18] Asmaa S, Daren F, Elisa B, *et al.* Data and syntax centric anomaly detection for relational databases [J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2016, 6 (6): 231-239.

[19] Han Jiawei, Kamber M, Pei Jian, 等. 数据挖掘：概念与技术 [M].

范明，孟小峰，译．3 版．北京：机械工业出版社，(Han Jiawei, Kamber M, Pei Jian, et al. Data mining: concepts and techniques[M].Fan Ming,Meng Xiaofeng,Translated. 3rd ed. Beijing: Mechanical Industry Press, 2012: 293-294.)

[20] 李航. 统计学习方法 [M]. 北京：清华大学出版社, 2012: 47-53. (Li Hang. Statistical learning method [M]. Beijing: Tsinghua University Press, 2012: 47-53.)

[21] Islam S M, Kuzu M, Kantarcioglu M. A dynamic approach to detect anomalous queries on relational databases [C]// Proc of the 5th ACM Conference on Data and Application Security and Privacy. San Antonio: Association for Computing Machinery, 2015: 245-252.

[22] Ronao A C, Cho S B. Mining SQL queries to detect anomalous database access using Random Forest and PCA [C]// Proc of the 28th International Conference on Current Approaches in Applied Artificial Intelligence. Berlin: Springer-Verlag, 2015: 151-160.

[23] TPC Benchmark C Standard specification revision 5. 11 [EB/OL]. (2010-01) [2018-08-23] http://www. tpc. org/tpc_documents _cu-rrent_versions/pdf/tpc-c_v5. 11. 0. pdf.